LA-UR-80-1542

CONF-800826--1

TITLE: PERCENTILE ESTIMATION USING THE NORMAL AND LOGNORMAL PROBABILITY DISTRIBUTION

AUTHOR(S): Thomas R. Bement

MASTER

SUBMITTED TO: American Statistical Meeting

University of California

LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545
An Affirmative Action/Equal Opportunity Employer

Form No. 836 R3
St. No. 2629
12/78

UNITED STATES
DEPARTMENT OF ENERGY
CONTRACT W-7405-ENG 36

PERCENTILE ESTIMATION USING THE NORMAL AND LOGNORMAL PROBABILITY DISTRIBUTION

by

Thomas R. Bement
Los Alamos Scientific Laboratory

ABSTRACT

Implicitly or explicitly percentile estimation is an important aspect of the analysis of aerial radiometric survey data. Standard deviation maps are produced for quadrangles which are surveyed as part of the National Uranium Resource Evaluation. These maps show where variables differ from their mean values by more than one, two or three standard deviations. Data may or may not be log-transformed prior to analysis. These maps have specific percentile interpretations only when proper distributional assumptions are met. Monte Carlo results are presented in this paper which show the consequences of estimating percentiles by i) assuming normality when the data are really from a lognormal distribution, and ii) assuming lognormality when the data are really from a normal distribution.

INTRODUCTION

Many types of investigations in geology and other disciplines are related to the problem of percentile estimation. The problem that motivated this study is the interpretation of aerial radiometric data being collected by the U.S. Department of Energy (DOE) as part of the National Uranium Resource Evaluation (NURE) program. Since 1974 the Grand Junction, Colorado Office of DOE, through its contractors, has been conducting aerial surveys over various portions of the United States. The data collected include observations in the gamma-ray portion of the spectrum from which the contributions of potassium

(^{40}K), uranium (from ^{214}Bi) and thorium (from ^{208}Tl) to total activity can be determined. Implicitly or explicitly, percentile estimation has always been an important feature of the treatment of this data.

A traditional method of displaying the results of an aerial survey is the standard deviation map. An example using ^{214}Bi is shown in Fig. 1 (Ref. 1). Flight lines are plotted on a map which can be used to overlay a 1:250,000 National Topographic Map Service (NTMS) quadrangle. Data are typically analyzed on a within geologic type basis. Observations (referred to as records) which are one, two or three standard deviations above or below the mean are indicated by one, two or three points plotted above or below the flight line. In some cases a lognormal distribution is assumed and log-transformed data is used to produce maps or portions of maps.

Under the assumption of normality (or lognormality) standard deviation maps have obvious percentile interpretations. This is important because large values of high percentiles may indicate a potentially favorable area. The problem to be addressed here is that of determining the effect of incorrect distributional assumptions on percentile estimation. Examination of the data indicates that there are many more distributional possibilities than the normal and lognormal, but for the purposes of this report it will be assumed that they are the only alternatives. In particular, two problems will be considered. They are a) determining the consequences of assuming a lognormal distribution when the distribution is really normal and b) determining the consequences of assuming a normal distribution when the distribution is really lognormal. Since the most common practice in NURE data presentation is to simply compute the mean and standard deviation of the untransformed data without doing any goodness-of-fit tests, the error associated with the second problem is probably the most common.

COMPARISON OF TWO TYPES OF ERRORS

It can be shown that the difference between a "correct" and an "incorrect" percentile estimate, expressed as a percentage of the correct estimate, depends on the first two moments of the distribution (either normal or lognormal) only through the coefficient of variation. The cases to be considered are those where the coefficient of variation is between 0.143 (expected value equal to 7.0 times the standard deviation) and 0.4 (expected value equal to 2.5 times the standard deviation). For values of the coefficient of variation below this region, the normal and lognormal are enough alike so that the differences are not of practical interest. A value of the coefficient of variation above this range implies that the corresponding normal distribution has a significant probability of negative values and since the data with which we are dealing is strictly positive a normal should probably not be used.

We shall now consider the error incurred in percentile estimation by assuming that data from a (truncated) normal distribution are actually lognormal. Consider a random sample of size n from $X \sim N(\mu, \sigma^2)$. An unbiased estimator of $\mu + Z_\alpha \sigma$, the 100α percentile, is $\bar{X} + Z_\alpha (a_n S)$ where Z_α is the 100α percentile of the standard normal distribution, S is the square root of the bias corrected maximum likelihood estimator of σ^2 and a_n is a bias correction factor for the standard deviation (Ref. 2). The factor a_n approaches 1.0 as n increases and since sample sizes are typically large (usually greater than 50 and often greater than 1000) it will be assumed to be equal to 1.0. Thus it is assumed that the "correct" estimator for the 100α percentile is $\bar{X} + Z_\alpha S$. If we mistakenly take the data to be from a lognormal distribution the 100α percentile is estimated by $\exp(\bar{Y} + Z_\alpha S_Y)$ where $\bar{Y}$ and S_Y are the mean and standard deviation of the log-transformed data.

A Monte Carlo study was conductred to determine the magnitude of the difference between the two estimators. Ignoring any truncation and for fixed sample size, this difference, expressed as a percentage of the correct estimator depends on the percentile being estimated and the coefficient of variation. Using Kinderman and Ramage's (Ref. 3) generator on a Los Alamos Scientific Laboratory CDC 7600 computer, 900 samples for each of several sample sizes (see Table 1) were taken from a standard normal distribution. Uniforms required by this procedure were supplied by a multiplicative congruential random number generator. The underlying population mean of each sample was transformed to provide each of the coefficients of variation listed in Table 1. The sample sizes listed are the number of observations after truncating any zero or negative values.

Percentile estimates using both methods listed above were obtained for each of the 900 samples. The means and standard deviations of each of these estimates were computed over the 900 samples and the difference of the two means was expressed as a percentage of the correct one. Percentage differences computed in this manner and based on a sample size of 2000 were used in Fig. 2. Figure 2 shows the region where the error (i.e., the percentage difference between the average estimators) is greater than and less than 5 percent. The estimated standard deviation of the percentage difference values is less than 0.4 percent. This figure also shows the error lines associated with treating a sample from a lognormal distribution as if it were normal.

Now, consider the error incurred in percentile estimation by assuming that data from a lognormal distribution are normal. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a lognormal distribution having underlying normal parameters μ and σ^2. That is, if $y = \ln X$ then $Y \sim N(\mu, \sigma^2)$. Let X_α denote the true 100α percentile of the distribution. The maximum

likelihood estimator of X_α is $\exp(\bar{Y} + Z_\alpha S_Y)$ where $\bar{Y}$ and S_Y are the maximum likelihood estimators of μ and σ^2. In practice, within the setting of the problem that motivated this investigation, S_Y is usually taken to be the square root of the bias corrected maximum likelihood estimate of σ^2. For the following, it will be assumed that this is the case. If we mistakenly assume the data are from a normal distribution, X_α will be estimated by $\bar{X} + Z_\alpha S$ where $\bar{X}$ and S are the sample mean and standard deviation (assume a large sample size).

A Monte Carlo study was conducted to determine the magnitude of the error resulting from the incorrect assumption of normality. The same sample sizes and coefficients of variation were considered here as in the previous case. Nine hundred samples of each of the listed sizes were taken from a standard normal distribution. The underlying population mean and variance were transformed so that the exponential of the sampled random variable would have the desired coefficient of variation.

Percentiles were estimated for each of the 900 samples assuming a normal distribution. Differences of each estimate from the true percentile, expressed as a percentage of the true percentile, were computed. The mean and standard deviation of these estimates and differences were computed over the 900 samples. The average percentage differences, based on samples of size 2000, were used in Fig. 2. The estimated standard deviation of these average differences is less than 0.1 percent. As in the case of incorrectly assuming lognormality, Figure 2 shows the regions where the error resulting from a false assumption of normality is less than and greater than 5 percent.

CONCLUSIONS

Figure 2 shows that making an error by assuming the data are from a normal population when they really are from a lognormal population causes problems primarily when estimating tail percentiles. When estimating lower tail percentiles, the error of failing to perform a log-transformation has much more serious consequences than does the error of transforming. When estimating percentiles that are between the 10^{th} and 80^{th} the consequence of either type of error is not great if the coefficient of variation is reasonably small, say less than 0.27. When estimating upper percentiles, the error of making a log-transformation has much more serious consequences than does the error of failing to transform. Table 2 provides details concerning this fact. For example, an incorrect assumption of normality results in a 16 percent error when estimating the 99^{th} percentile of a distribution having coefficient of variation equal to 0.417. Under the same conditions, the error of incorrectly assuming that the distribution is lognormal results in a 66 percent error. A 66 percent error could be significant, as the following example illustrates.

A mean ^{214}Bi count rate of 30 counts per second (cps) is typical of many geologic formations in the Rawlins, Wyoming, survey (Ref. 1). Assuming a distribution is really normal with a mean of 30 and a coefficient of variation equal to 0.417, the 99^{th} percentile is equal to 59 cps. If lognormality were incorrectly assumed, a 66 percent error of 39 cps would result. In the Rawlins survey, a ^{214}Bi activity of 9.6 cps is equivalent to one part per million (ppm) uranium. Thus, an error of 39 cps amounts to a 4.1 ppm uranium error.

Throughout this report it has been assumed that the aerial radiometric data are either normally or lognormally distributed and that no other alternatives exist. As mentioned earlier, examination of the data indicates that other possibilities should be considered. Studies are now underway to evaluate several methods of estimating percentiles by comparing them over a broad range of distributions.

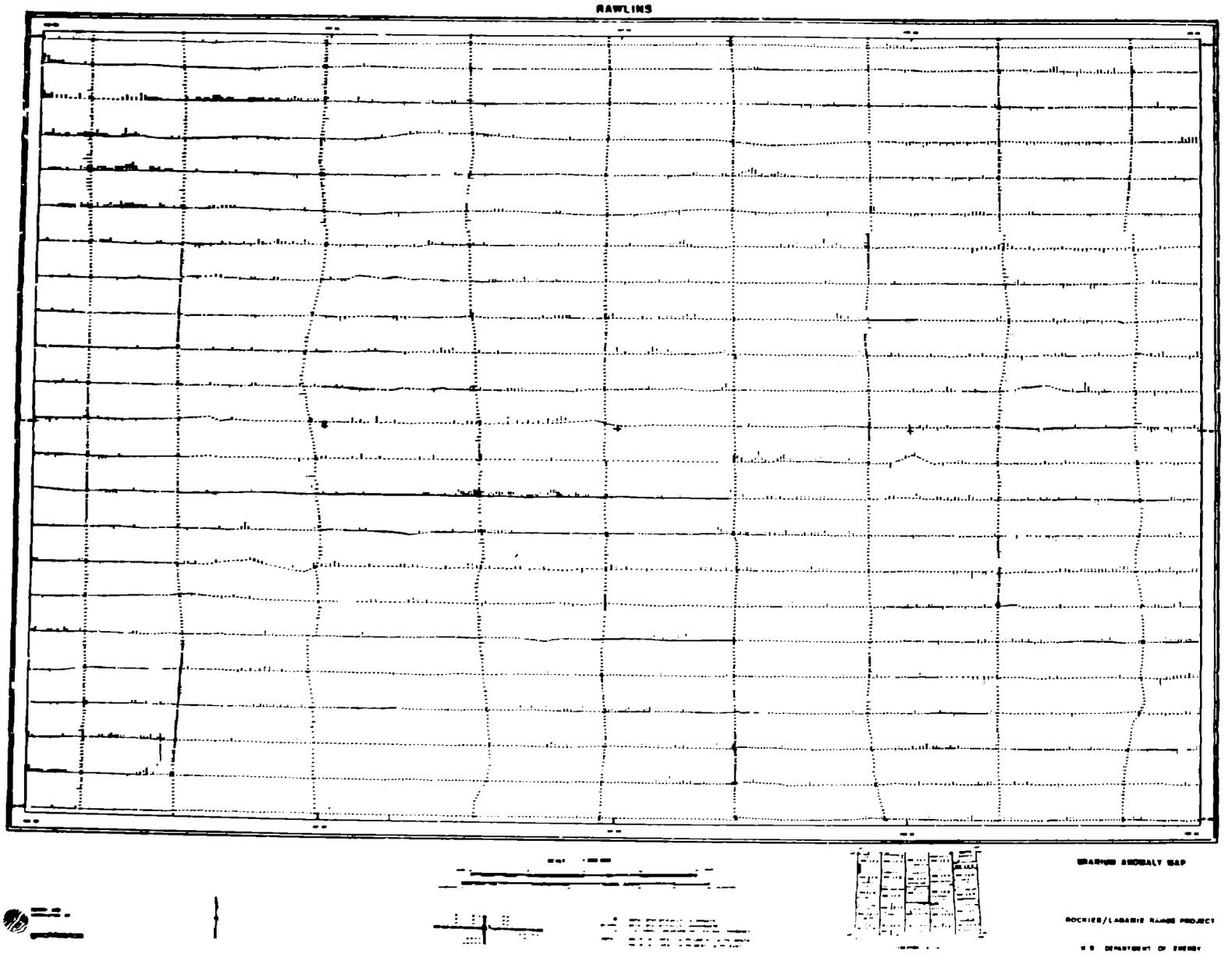

Fig. 1. An example of a standard deviation map prepared from aerial radiometric survey data (Ref. 1).

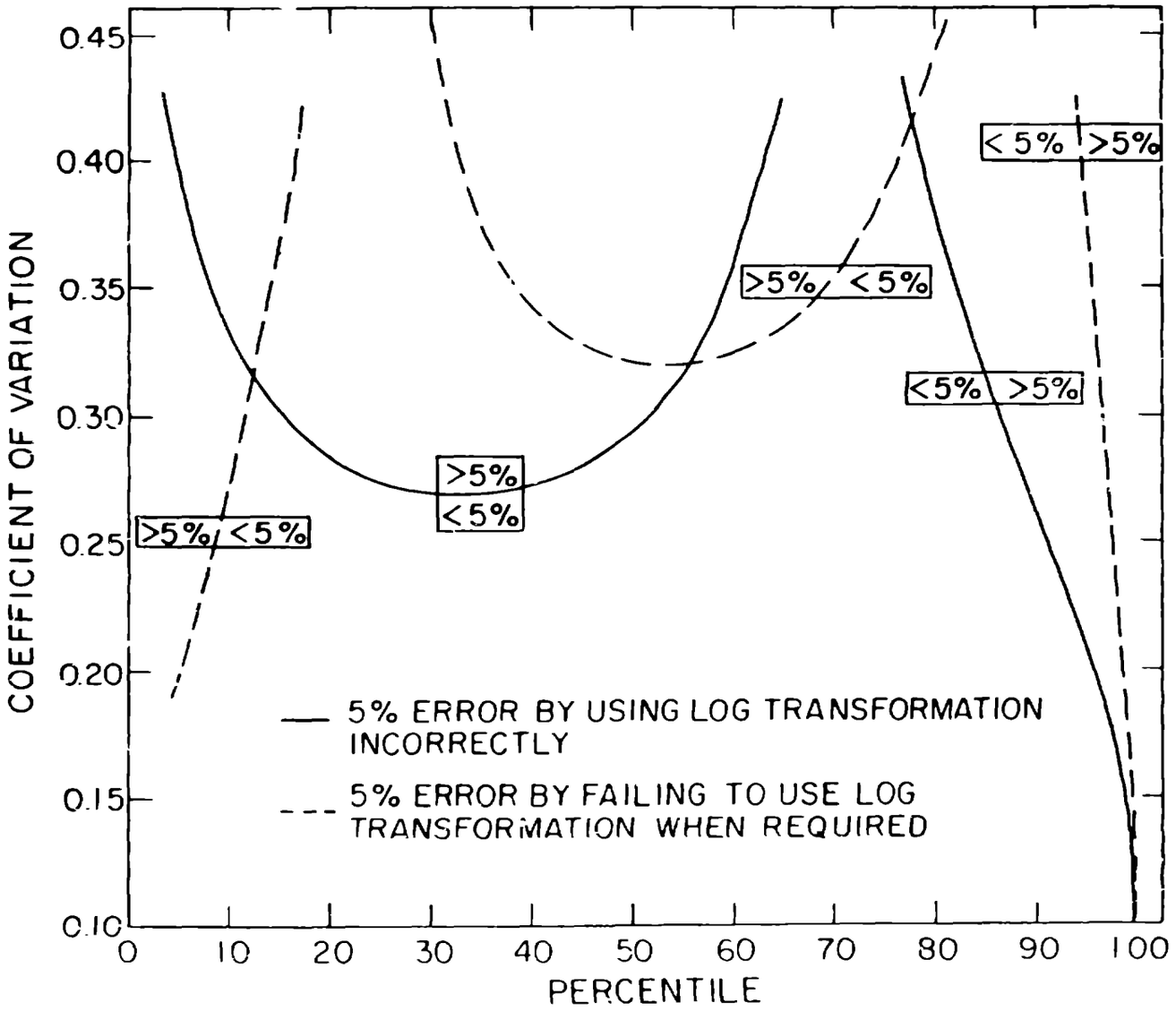

Fig. 2. Regions where incorrect choice of distribution leads to errors greater than and less than 5 percent.

Table 1

Percentiles, Coefficients of Variation and

Sample Sizes Used in Simulations

Percentiles	Coefficient of Variation	Sample Size
0.05	1/2.4 = 0.417	20
0.10	1/2.6 = 0.385	50
0.20	1/2.8 = 0.357	100
0.30	1/3.0 = 0.333	200
0.40	1/3.3 = 0.303	300
0.50	1/3.6 = 0.278	500
0.60	1/4.0 = 0.250	1000
0.70	1/4.5 = 0.222	2000
0.80	1/5.0 = 0.200	
0.90	1/5.5 = 0.182	
0.975	1/6.5 = 0.154	
0.99	1/7.0 = 0.143	
0.995		
0.999		

All combinations were considered.

Table 2

Comparison of Two Types of Errors when Estimating Upper Percentiles

Coefficient of Variation	95th Percentile Percentage Error by Incorrectly Assuming Normality	95th Percentile Percentage Error by Incorrectly Assuming Lognormality	99th Percentile Percentage Error by Incorrectly Assuming Normality	99th Percentile Percentage Error by Incorrectly Assuming Lognormality	99.9th Percentile Percentage Error by Incorrectly Assuming Normality	99.9th Percentile Percentage Error by Incorrectly Assuming Lognormality
.250	< 5	8	8	16	15	28
.278	< 5	11	9	22	17	38
.303	< 5	14	11	28	19	49
.333	< 5	19	13	37	22	65
.357	< 5	23	13	45	23	79
.385	< 5	27	14	54	26	96
.417	6	33	16	66	28	117

REFERENCES

1. Aerial Gamma Ray and Magnetic Survey, Rock Springs, Rawlins and Cheyenne Quadrangles, Wyoming and the Greeley Quadrangle, Colorado. Final Report (prepared for the Department of Energy, Grand Junction Office, Grand Junction, Colorado, by GeoMetrics, Sunnyvale, California, December 1980.

2. N. L. Johnson and S. Kotz, Continuous Univariate Distributions - 1, Houghton Mifflin Co., New York, 1970.

3. A. J. Kinderman and J. G. Ramage, "Computer Generation of Normal Random Variables," Journal of the American Statistical Association, 71, 893-896, 1976.